Worry about misuse of AI, not superintelligence

Arvind Narayanan and Sayash Kapoor
Aug 8, 2024

CEOs of AI companies have said that we should worry about AI becoming "superintelligent", some even arguing it could lead to humanity's demise. OpenAI CEO Sam Altman expects AGI, or artificial general intelligence—AI that outperforms humans at most tasks—around 2027 or 2028. Elon Musk's prediction is 2025 or 2026, and he has claimed that he was "losing sleep over the threat of AI danger."

But predictions of imminent human-level AI have been made for over fifty years, since the earliest days of AI. Those predictions were wrong — not merely in terms of timing but in a deeper, categorical sense. We now know that the specific approaches to AI that experts had in mind in the past are simply incapable of getting to human-level intelligence. Similarly, as the limitations of current AI have become clear, most AI researchers have come around to the view that simply building bigger and more powerful chatbots won't lead to AGI.

Advanced AI is certainly a long-term worry that researchers should study. But AI risks in 2025 will arise from misuse, as they have thus far. These might be unintentional misuses, such as lawyers over-relying on AI. After the release of ChatGPT, many lawyers have been sanctioned for using AI to generate erroneous court briefings, apparently unaware of chatbots' tendency to make stuff up. In British Columbia, Chong Ke was ordered to pay costs for opposing counsel for including fictitious AI-generated cases in a legal filing. In New York, Steven Schwartz and Peter LoDuca were fined $5,000 for providing incorrect citations. In Colorado, Zachariah C. Crabill was suspended for a year for using fictitious court cases generated using ChatGPT and blaming a "legal intern" for the mistakes. The list is growing quickly.

Other misuses are intentional. In January 2024, sexually explicit deepfakes of Taylor Swift flooded social media platforms. These images were created using Microsoft's AI tool "Designer". While the company had guardrails to avoid generating images of real people, misspelling Swift's name was enough to get around this. Microsoft has since fixed this error. But Taylor Swift is the tip of the iceberg, and non-consensual deepfakes are proliferating widely — in part because open-source tools to create deepfakes are available publicly. Ongoing legislation across the world seeks to combat deepfakes in hope of curbing the damage. Whether it is effective remains to be seen.

In 2025, it will get even harder to distinguish what's real from what's made up. The fidelity of AI-generated audio, text, and images is remarkable, and video may be next. This could lead to the "liar's dividend": those in positions of power repudiating evidence of their misbehavior by claiming that it is fake. In 2023, Tesla argued that a 2016 video of Elon Musk could have been a deepfake in response to allegations that the CEO had exaggerated the safety of Tesla autopilot leading to an accident. An Indian politician claimed that audio clips of him acknowledging

corruption in his political party were doctored (the audio in at least one of his clips was verified as real by a press outlet). And two defendants in the January 6 riots claimed that videos they appeared in were deepfakes. Both were found guilty.

Meanwhile, companies are exploiting public confusion to sell fundamentally dubious products by labeling them “AI”. This can go badly wrong when such tools are used to classify people and make consequential decisions about them. Hiring company Retorio claims that its AI predicts candidates' job suitability based on video interviews, but a study found that the system can be tricked simply by the presence of glasses or by replacing a plain background with a bookshelf, showing that it relies on superficial correlations. The Netherlands used an algorithm to identify people who committed child welfare fraud. It wrongly accused thousands of parents, often demanding to pay back tens of thousands of euros. The system was in place for six years before it was uncovered. In the fallout, the Prime Minister and his entire cabinet resigned. There are dozens of applications in healthcare, education, finance, criminal justice, and insurance where AI is used to deny people important life opportunities.

In 2025, we expect AI risks to arise not from AI acting on its own but because of what people do with it. That includes cases where it *seems* to work well and is over relied upon (lawyers using ChatGPT), when it works well and is misused (non-consensual deepfakes and the liar's dividend), and when it is simply not fit for purpose (denying people their rights). Mitigating these risks is a mammoth task for companies, governments, and society. It will be hard enough without getting distracted by sci-fi worries.